A multifaceted approach to investigating pre-task planning effects on paired oral test performance

Ryo Nitta *Nagoya Gakuin University, Japan* and

Fumiyo Nakatsuhara *University of Bedfordshire, UK*

Abstract

Despite the growing popularity of paired format speaking assessments, the effects of pre-task planning time on performance in these formats are not yet well understood. For example, some studies have revealed the benefits of planning but others have not. Using a multifaceted approach including analysis to extend the process of performance, the aim of this paper is to investigate the effect of pre-task planning in a paired format. Data were collected from 32 students who carried out two decision-making tasks in pairs, under planned and unplanned conditions. The study used analyses of rating scores, discourse analytic measures, and conversation analysis (CA) of test-taker discourse to gain insight into co-constructing processes. A post-test questionnaire was also administered to understand the participants' perceptions toward planned and unplanned interactions. The results from rating scores and discourse analytic measures revealed that planning had limited effect on performance, and analysis of the questionnaires did not indicate clear differences between the two conditions. CA, however, identified the possibility of a contrastive mode of discourse under the two planning conditions, raising concerns that planning might actually deprive test-takers of the chance to demonstrate their abilities to interact collaboratively.

Key words

pre-task planning, paired oral tests, task-based language teaching, conversation analysis, interactional competence, co-construction

I Introduction

Task-based language teaching (TBLT) and language testing research have a reciprocal relationship in that issues proposed by task-based researchers have been applied and investigated in testing contexts and the results have then been fed back into task-based research (e.g., Elder, et al. 2002; Iwashita, et al., 2001; Skehan, 1998). One actively researched area within such a task-test cycle has been on the effect of pre-task planning on L2 oral performance.

Planning time prior to a task is considered beneficial in the area of TBLT from a cognitive perspective. Limited working memory capacity makes it difficult for learners to focus attention on formal aspects of production during performance. Given planning time, learners could prioritize meaning while retaining focus on form on-task. In addition, planning time is likely to encourage learners to access explicit (analytic) knowledge, as they have limited implicit (automatized) knowledge that can effortlessly be accessed on-task. Under these cognitive principles, task-based researchers have investigated how methods of pre-task planning influence oral performance (e.g., different lengths of planning time in Mehnert, 1998; unguided/guided planning in Foster & Skehan, 1996). Findings have varied depending on the nature of planning, task types, and proficiency levels of learners, but a general consensus by these researchers is that relatively long planning times (e.g., 10 minutes) in classroom and laboratory settings provide clear benefits to task performance in terms of fluency, but to a lesser extent to complexity and accuracy (see Ellis, 2009, for a comprehensive review).

The role and value of pre-task planning time is also an issue of relevance in language testing research and practice. Pre-task planning has been operationalized in a number of large-scale standardized speaking tests such as IELTS and TOEFL, particularly prior to monologic tasks. The provision of pre-task planning time has been discussed as a way to establish a fair environment for test-takers in these tasks, by recognizing that planning time helps to control the level of cognitive demand imposed by potentially unfamiliar topics and enabling test-takers to produce their best possible performance (Field, 2011).

Methods of delivering pre-task planning in testing research and practice are rather uniform, i.e., always using unguided planning for relatively short periods (e.g., 1 minute in Wigglesworth, 1997, 3 minutes in Elder & Iwashita, 2005, and 5 minutes in Wigglesworth, 2001); however, findings in testing research have been mixed. While positive effects were found by Wigglesworth (1997), Tavakoli and Skehan (2005), and Xi (2005), limited benefits were reported by Wigglesworth (2001), Elder and Iwashita (2005), and Wigglesworth and Elder (2010). One possibility for limited effects of planning on test performance may be related to the high stakes contexts of language testing. That is, as a testing context is likely to lead to increased attention by test-takers in regard to the accuracy of their output language, resulting in careful "on-line" planning while they are speaking, potentially beneficial effects of pre-task planning may be over-ridden (Ellis, 2005).

Differences in effects associated with the provision of pre-task planning might also result from different methods of analysis, i.e., discourse analytic performance measures (e.g., fluency, complexity, and accuracy) applied in task-based research and raters' assessments applied in testing research. Discourse analytic measures seem to more

sensitively register subtle differences caused by planning when compared to raters' judgements. Wigglesworth (1997) observed that even trained raters, who rely on impressionistic judgments guided by generic descriptions, were unable to make the fine distinctions yielded by discourse analytic measures.

1 Pre-Task Planning in Dialogic Tasks

In addition to the possible influence of the contexts of investigation, task formats could mitigate or exaggerate the effects of planning. One neglected area is how planning time influences performance in dialogic tests. Thus far, testing studies on pre-task planning have exclusively used monologic types of task (e.g., picture descriptions and monologues on given topics).

Clear differences between monologic and dialogic tasks can be considered in terms of performance processes. Once a task starts in a solo performance, the speaker can only rely on his/her own resources. The speaker needs to find a solution for him/herself to continue speaking and to construct the whole performance. In contrast, the whole conversation in dialogic tasks is co-constructed as a consequence of iterative language exchange processes. The conversational path is continuously open and subject to utterances by both parties.

Much attention in speaking assessment practice and research has recently been focused on the co-constructing process in interactive tasks. Paired and group speaking formats are now widely utilized in both high- and low-stakes tests to assess test-takers' communicative abilities including initiating and maintaining interactions (e.g., Cambridge ESOL Main Suite examinations, the College English Test-Spoken English Test in China, and the Hong

Kong Advanced/Supplementary Level Examination). Recent speaking test validation models such as Weir's (2005) socio-cognitive framework include the interlocutors' input language as one of the contextual parameters that could influence test takers' performance in speaking tests.

A number of studies have demonstrated different ways used by test-takers in co-constructing interactions with different types of tasks and conditions under which tasks are implemented (e.g., Galaczi, 2008; Nakatsuhara, 2011; Van Moere, 2007). For example, Galaczi (2008) identified three distinct global patterns for interactions in the paired discussion part of the Cambridge FCE, viz., *collaborative, parallel*, and *asymmetric*. In the *collaborative* pattern of interactions, participants would shift their interactional roles between listener and speaker, and support the development of both topics. The *parallel* pattern resembled "solo versus solo" interaction, in which both speakers would initiate and develop their own topics but would have limited engagement with the other's ideas. The *asymmetric* pattern was characterized by unbalanced contributions to the quantity of talk and topic development in the dyad, with one speaker leading the interaction and the other taking a secondary role. Galaczi (2008) revealed that high scores on the "interactive communication" scale were generally associated with a collaborative pattern of interaction, while a parallel pattern led to low scores. This has also been confirmed by Gan's (2010) group oral study.

Given the growing popularity of interactive formats in language testing, understanding the role and value of planning time should attract more attention in research. In fairness, a number of TBLT studies have investigated pre-task planning effects on

dialogic tasks. According to Ellis's (2010) review of pre-task planning literature, thirteen TBLT studies involved tasks in dialogic mode (e.g. including paired speakers), and demonstrated the benefits of pre-task planning to learner performance in such tasks. However, these dialogic studies analyzed paired performances collectively, and none of them discussed or investigated individual performances, especially how one individual's performance might interact with that of the partner. The gap in these research studies is probably due to the fact that they were primarily concerned with the cognitive complexity of different tasks (e.g., personal information, narrative, and decision-making tasks in Foster and Skehan, 1996) along with the linguistic demands of the task designs without taking the co-constructing aspects of interaction into account. Paired speaking tests, however, are intentionally designed to measure test-takers' *interactional competence* (Young, 2000) in addition to other linguistic aspects of performance. The additional concern with interactional competence means that it is highly important to understand whether/how provision of pre-task planning time influences interactive patterns of dialogue (Galaczi, 2008). This is because test-taker interaction affects test validity in important ways.

Therefore, given the emphasis on and value of co-constructed aspects of dialogic interactions and the popular use of paired formats in language testing, a closer look at the effects of planning time in interactive formats is appropriate. This should provide information that test designers can draw on to ensure their decisions are better informed when implementing pre-task planning.

2 Multifaceted approach

Most previous studies of planning time effects on L2 oral performance have mainly focused on aggregated outcomes for groups of learners that were aimed at finding significant differences across performance conditions. This line of approach was aimed at obtaining snapshots or summative views of learners. In particular, these summative studies usually abandoned detailed analysis of sequences of discourse moves through which learners dealt with tasks, assuming that their performance was essentially the same throughout the task (Samuda & Bygate, 2008). It is however quite possible that performance at the beginning could differ markedly from performance during other phases, as reported by Skehan and Foster (2005). We agree that performance is not a simple sum of single productions, in the recognition that interactions involve a non-linear process through iterative turn-taking opportunities. We intend to reconcile the traditional summative approach with one that is more process-oriented to gain insight into similarities and differences in the processes of learners' interactions under different planning conditions.

In addition to understanding interactional processes, it would be worth investigating test-takers' perceptions toward pre-task planning. Wigglesworth and Elder (2010) argued that 1 minute of planning would be important to enhance the face validity of tests from their questionnaire and interview analyses using monologic types of task. Weir et al. (2006) developed a cognitive processing questionnaire to explore how test-takers responded to planned and unplanned monologic tests and found what test-takers thought or did during the planning stage and while they were performing tasks. As there have been no such studies with dialogic test formats, we investigated test-takers' perceptions toward planned and unplanned dialogic performances.

This study addresses four questions:

- RQ1: Does pre-task planning affect test-takers' performance in paired oral interactions as measured by rating scores?
- RQ2: Does pre-task planning affect their language performance as measured by discourse analytic measures?
- RQ3: How do test-takers perceive the usefulness of pre-task planning time, and their own performance under planned and unplanned conditions?
- RQ4: How do test-takers co-construct paired oral performance under planned and unplanned conditions?

The first two questions are concerned with traditional approaches to researching pre-task planning effects. The third and fourth questions are aimed at explicating and elaborating on the statistical findings obtained from the first two questions. The third question brings in test-takers' opinions about their feelings and perceptions about the issue in focus, and the last question explores the co-constructing processes of paired performances by taking a process-oriented approach.

II Method

1 Participants

Thirty-two English majors at a Japanese university participated in this study. They were either in their second or third year, and the average length of their English studies was 8.52

years (SD=1.43). Their first language was Japanese, and gender was balanced (males: N=16, females: N=16). None of them had notable experience of living in an English-speaking country. Their English proficiency-level was considered to be around B1 (Threshold) of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), judging from their recent TOEFL IPT scores (involving listening, reading, and grammar) with an average score of 476.41 (SD= 31.24; see ETS, 2012). Class teachers also confirmed that their oral proficiency was around the B1 level with a few exceptions at the B2 level. This group was thus relatively homogeneous in terms of their age, first language, educational background, and English proficiency. They were freely paired with their friends, and all 16 pairs were same gender pairs to control for other possible confounding variables (acquaintanceships and gender; see O'Sullivan, 2008). All the test-takers were preparing to study abroad when the data were being collected, and a speaking test was administered as part of a pre-departure assessment of English language abilities.

2 Design

The 16 pairs took a speaking test consisting of one warm-up task and two decision-making tasks under two different conditions with a three-minute pre-task planning time (+) and without a planning time (–). The order of the + and – planning conditions and task prompts were counterbalanced across the 16 pairs (see Table 1 for the first eight pairs) to balance the practice effect across the two performances. They performed each task for five minutes. All test sessions were video-recorded. Immediately after they had performed all tasks, they

completed a questionnaire on their perceptions of the tasks and the planning time that was provided.

Table 1. Task sequence

Pair	1st Task	2nd Task	3rd Task
1	Warm-up	Prompt A+	Prompt B–
2		Prompt A–	Prompt B+
3		Prompt B+	Prompt C–
4		Prompt B–	Prompt C+
5		Prompt C+	Prompt D–
6		Prompt C–	Prompt D+
7		Prompt D+	Prompt A–
8		Prompt D–	Prompt A+

Note: + with planning time, – without planning time

3 Tasks

The warm-up task was first presented for two minutes, in which we asked them to introduce each other, and this was followed by two decision-making tasks. The decision-making tasks were adapted from the "Part 3 collaborative task" from the Cambridge First Certificate in English (FCE) speaking test, which was aimed at assessing learners' interactional abilities including *sustaining an interaction, exchanging ideas, agreeing and/or disagreeing, suggesting*, and *reaching a decision through negotiation* (Cambridge ESOL, 2012; see also Taylor, 2011, for useful information about the design and nature of the FCE speaking tasks, together with context and cognitive validity evidence to support claims that they are at B2 level). The target level of FCE is CEFR B2 and thus the tasks were considered to be a little too difficult for most participants in this study. However, the decision was made to use FCE tasks rather than easier tasks, e.g., from the Cambridge Preliminary English Test (PET).

This is because the present study was part of a larger study that compared participants' levels of proficiency before and after one-year of study experience abroad, and therefore the test needed to be relevant to assess progress they would make in a year. Furthermore, the selection of topics in FCE seemed more cognitively appropriate to the participating university students than those in PET.

Participants in the decision-making task were given both oral and written instructions with a prompt card with 7–8 visual items. They were first required to discuss each visual item in relation to the given topic (e.g., how important each item was for a happy life), and then asked to reach consensus on one or two items (e.g., which two items were the most important; see Cambridge ESOL, 2012, for an example task). Four different task prompts (A: Happiness, B: Profession, C: Café, and D: Tourists) were prepared, and two prompts were selected from the pool of four for each pair (see Table 1). Since the task prompts were taken from official Cambridge FCE practice papers (Cambridge ESOL, 2008), which included past FCE items, the difficulty of these prompts was calibrated to be comparable. In order to examine to what extent cultural aspects of the tasks were familiar to the participants, we asked their class teachers about the content and format of the four tasks and confirmed that these tasks should not cause particular difficulty for understanding. In addition, all participants were familiarized with this type of task during their preparatory course to study abroad, which was delivered by one of the researchers.

Although Cambridge FCE provides only 3 minutes to perform the task, we decided to extend it to 5 minutes in the present test. This was to elicit speech samples from both parties that could be rated, as we needed to award scores for this task only, unlike the real

FCE, where scores are awarded for the four tasks together.

4 Planning

A 3-min planning time was established for the present study because we thought a planning time of over 3 minutes would not be feasible in most testing contexts, while previous studies suggested that 1 minute might be too short for planning to have any effect on performance (e.g., Wigglesworth & Elder, 2010). The planning was unguided, and speakers under planned conditions were simply told to use their time to prepare to speak in any way they wished. They were instructed to plan individually and not to discuss plans with their partner before the test began. They were not allowed to use external resources (e.g., dictionaries or the Internet) but could take notes while planning, and they were informed they would be able to keep their notes while speaking to their partner. These instructions about pre-task planning were delivered both orally and in writing in English. The pre-task planning session was also videotaped, and we confirmed that all participants strictly followed these planning instructions.

5 Analysis

a. Rating Scores. The video-recorded performances of the 32 participants under the two conditions were rated using a modified version of the rating scale developed by Iwashita et al. (2001), which consisted of fluency, complexity, and accuracy (Appendix 1). Since the participants' levels of proficiency were expected to be clustered toward the bottom of the original scale especially with the FCE tasks we employed, the scale was modified by

adding Level 0 and a middle point between the levels to distinguish the participants more effectively. The usefulness and reliability of the modified scales will be reported in the Results section.

The ratings were carried out by two raters, both of whom held Ph.D.s in Applied Linguistics and who had extensive experience in language testing research and practice. Approximately 1.5 hours of rater training were provided, where they discussed rating descriptors and independently rated three video-recorded paired test sessions. After they had rated each session, the scores they awarded and reasons for each rating were discussed to achieve agreement. Then, all 32 video-recorded test sessions (16 sessions each under planned and unplanned conditions) were independently rated. The video clips were mixed, so that the rating of a pair under one condition would not affect the rating of the same pair under the other condition. They carried out the ratings in a counter-balanced manner; one rater started ratings from Video 1, while the other started ratings from Video 32. The scores were statistically analyzed using multi-faceted Rasch analysis with the FACETS program.

b Discourse analytic measures. The principal dimensions of the multi-componential nature of L2 performance and proficiency have been considered to be captured by the notions of fluency, complexity, and accuracy in the area of TBLT.

Researchers applied various measures of fluency, and these can be categorized into three subcategories of speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). Speed fluency refers to how fast the produced language is in terms of time units. Both breakdown and repair fluency aim at capturing dysfluent features of L2 oral production, but

breakdown fluency, as measured by the amount of silence, more directly concerns perceived feelings of dysfluency, while repair fluency is not necessarily a sign of dysfluency.

Complexity can be measured in terms of various dimensions to quantify the elaboration of language. This study applied the most frequently used syntactic complexity measure of oral performance, i.e., clauses per AS-unit. The AS-unit is an utterance consisting of an independent clause together with any subordinate clauses associated with it (Foster et al., 2000). In addition, the study used the Measure of Textual Lexical Diversity (MTLD: McCarthy & Jarvis, 2010) to analyze lexical complexity. MTLD was chosen from other lexical diversity measures such as type-token ratios, because this measure does not get distorted by text length.

A global measure of accuracy was applied in the present study rather than classifying types of linguistic errors or ranking the effects of inaccuracies. Although the percentage of error-free clauses has often been used as a global measure (e.g., Foster & Skehan, 1996), this ignores cases where there is more than one error in a clause, leading to possible biases. Thus, we counted the number of errors per 100 words (e.g., Mehnert, 1998), which seems more sensitive to the proportion of accurate production, as it takes into account all the errors produced. All errors related to syntax, morphology, and lexical choice were considered while those related to phonology and discourse (e.g., communicative effectiveness) were not considered. The definition of ‘errors’ here was also applied to the accuracy rating scale (Appendix 1).

As the present study focused on the impact of planning time on dialogic test performance, it was necessary to include measures that were sensitive to interactional differences created by planning time. The number of words per turn (Duff, 1986) was counted for this purpose. Higher values for this were indicative of a long speech with less interaction while lower values would suggest more frequent turn-taking.

All paired speeches were fully transcribed and individual performances were coded for fluency, complexity, accuracy, and interaction as follows:

Fluency

- Speed: The total number of words per second (excluding pauses between turns) divided by the total length of speaking time
- Breakdown: The total length of pauses (longer than 0.2 second, including both intra-turn pauses and pauses between turns) divided by the total length of speaking time
- Repair: The number of repetitions, self-corrections, and reformulations, divided by the total number of words

Complexity

- Syntactic: Clauses per AS-unit
- Lexical: Lexical diversity (MTLD)

Accuracy

- Global accuracy: The number of errors per 100 words

Interaction

- Turn-length: The total number of produced words divided by the number of turns

After agreement was reached on the coding schemes and more specific guidelines (e.g., contractions such as “it’s” and “doesn’t” were treated as one word), all samples were examined by the two researchers separately to identify fluency markers, AS-units, and errors. All coded transcripts were then compared, and discrepancies were discussed and agreement was reached for every single case. Here, it should be noted that pauses between

turns were treated differently in measuring the speed and breakdown of fluency. Pauses between turns were excluded from the data for the speed of fluency, so that the speed of fluency would only account for the articulation rate and intra-turn pauses. In contrast, the breakdown of fluency included pauses between turns in the analysis as well as intra-turn pauses. It was not possible or even advantageous to determine the ownership of unfilled pauses between turns in dialogues where fluency was co-constructed as in a *confluence* (McCarthy, 2010). Both conversants were responsible for such pauses unless the previous speaker had explicitly nominated the next speaker (e.g., by questioning). Therefore, these pauses whose ownership was unidentifiable were divided by two, and half the pause duration was added to both conversants.

c Questionnaire data. Weir et al.'s (2006) cognitive processing questionnaire was used with some modifications (see Appendix 2). The questionnaire consisted of four parts: about the tasks (Part 1), about what the participants thought of or did before they started (Part 2), about the planning stage (Part 3), and about what occurred while they were speaking (Part 4). Two items (Q13–14) in Part 3 and three items (Q4–6) in Part 4 were added to the original version to reflect unique features in paired interactions. The results from Parts 2 to 4 are reported following the classification of items based on the findings in Weir et al.'s (2006) factor analysis. Non-parametric tests were used for inferential statistics since the questionnaire data were all ordinal.

d Conversation Analysis. CA was carried out to investigate similarities and differences in the test-takers' co-constructing processes. A number of studies over the last two decades have analyzed the discourses of various speaking test formats using CA, and this has been recognized as an invaluable methodology to describe and validate oral proficiency tests (e.g., Brown, 2003; Lazaraton, 2002). Building on such contributions of CA, this study utilized CA methodology to interpret and elaborate on the findings revealed by the statistical analysis of rating scores and discourse analytic measures.

The recorded data were transcribed by the two researchers using a slightly simplified version of CA notation (Atkinson & Heritage, 1984; Appendix 3). As they transcribed, transcripts completed by one of the researchers were checked by the other to confirm their accuracy. All the modifications suggested were discussed, and the agreed rules were further applied to the rest of the transcriptions.

III Results

1 Score Analysis

The facet map in Figure 1 represents an overview of the rating results and plots examinee abilities, rater severity, and the difficulty of planning conditions and difficulty of rating categories, which are the four major sources (i.e., facets) for test score variance. They were all measured in the uniform units (*logits*) indicated on the left of the map (*measure*). More competent examinees were placed toward the top and less competent toward the bottom. More severe raters, and more difficult planning conditions and rating categories appear

toward the top, and more lenient raters and easier planning conditions and categories appear toward the bottom. The *scale* column refers to the rating scale steps used in this study. This map indicates that unplanned conditions were more difficult than those that were planned, and the category for accuracy was more difficult than those for complexity and fluency.

Add Figure 1 around here

As expected, the test was generally difficult for the participants. However, this did not seem to distort or degrade the measurement system. The fit statistics of all facets measured with infit mean square values indicated that the indices ranged from 0.7 to 1.3, being well within an acceptable range of 0.5 to 1.5 (Wright & Linacre, 1994). This suggests that the response of participants, raters, and planning conditions, and three rating categories were all "productive for measurement" (ibid.). The separation index was 4.40, and the examinees were able to be separated into 6.20 statistically separate strata, although we need to bear in mind that participants' scores were clustered toward the bottom of the rating scale. The person reliability of 0.95 was also acceptable. The rater reliability we obtained was the best one could obtain from FACETS (i.e., 0.0), suggesting that the two raters were interchangeable (Myford & Wolfe, 2004). This overall analysis, therefore, indicates that the modifications we made to the original rating scales did not cause problematic inconsistencies and that the rating of speech samples in this study was reliably carried out.

As illustrated in Figure 1, the planned conditions were easier than those that were unplanned. Table 2 provides more details on these differences. The FACETS program yields several statistical measures for the differences between the elements of each facet.

One such measure is the fixed chi-square, which tests the null hypothesis that all elements of the facets are equal. The chi-square statistics revealed that the difference between the two conditions was statistically significant (χ^2=21.0, p<0.01), although the actual difference in scores indicated by the fair average scores was rather small (0.24).

Table 2. Rating category measurement report

Planning condition	Fair Average	Measure	Infit MnSq	Fixed (all same) chi-square
+	2.52	-.36	.89	χ^2=21.0, p<.01
-	2.28	.36	1.03	

Each rating category was then analyzed for any impact of planning time. As listed in Table 3, the analysis of each rating category indicated that planning made a statistically significant difference to fluency (χ^2=17.7, p<0.01) and complexity (χ^2=5.8, p=0.02), and the p value for accuracy also approached significance (χ^2=4.0, p=0.05). While the differences in fair average scores were small (i.e., only 0.44 for fluency, 0.18 for complexity, and 0.12 for accuracy), there was a trend where the test-takers performed slightly better under the planned conditions. The difference for fluency was the most marked of these differences in scores.

Table 3. Impact of planning condition on each rating category

Rating Category	Planning Condition	Fair Average	Measure (difficulty)	Infit MnSq	Fixed (all same) chi-square
Fluency	+	2.90	-.70	.79	χ^2=17.7
	-	2.46	.70	1.16	p<.01
Complexity	+	2.46	-.32	.79	χ^2=5.8
	-	2.28	.32	1.09	p=.02
Accuracy	+	2.30	-.31	.94	χ^2=4.0

	-	2.18	.31	.94	p=.05

2 Discourse analytic measures

The second research question concerned the effects of pre-task planning on performance quantified by various discourse analytic measures. Table 4 presents the descriptive statistics and impacts of pre-task planning using a paired t-test. Planning time tended to lead to significantly fewer numbers of words per second (speed fluency) and pauses (breakdown fluency), whereas planning did not result in any differences in terms of complexity or accuracy. The most remarkable characteristic was found in interaction. Close examination of each pair in the turn-length revealed that longer utterances were produced under the planned conditions in most of the pairs (14 out of 16 pairs), although the average turn-lengths varied among the pairs. The results obtained here will be further discussed in conjunction with the conversation analysis below.

Table 4. Impact of planning condition on discourse analytic measures

Focus	**Measure**	**Planning condition**	**Mean**	**S.D.**	**Paired samples t-test**
Speed fluency	Number of words per second	+	1.14	0.34	t=-.093
		-	1.34	0.37	p<.001
Breakdown fluency	Length of pauses per second	+	0.63	0.19	t=-.202
		-	0.72	0.18	p<.001
Repair fluency	Number of dysfluent features per session	+	16.00	8.52	t=.040
		-	16.06	8.69	p=.968
Syntactic complexity	Number of Clauses per AS-unit	+	1.11	0.11	t=-1.741
		-	1.07	0.07	p=.092
Lexical complexity	Lexical diversity	+	26.71	7.04	t=-.586
		-	27.53	8.32	p=.562
		-	7.97	7.19	

Accuracy	Number of Errors per 100 words	+	7.60	2.79	t=.093
		-	7.67	4.03	*p*=.927
Interaction	Number of words per turn	+	10.02	7.23	t=2.743
		-	7.97	7.19	*p*=.010

3 Questionnaire

The questionnaire aimed to identify any tendencies of test-taker perceptions toward the tasks, planning, and performances. The findings in Part 1 (about tasks) indicated that participants perceived the language and information used in the four prompts as comparable, in terms of their lexical and syntactic difficulty, information abstractness, and topic familiarity. They also reported that preparation time and task time were more or less appropriate. No significant differences were detected in these responses either across the four prompts or between planned and unplanned conditions (see Tables 5 & 6 in Appendix 2).

The means for the responses in Part 2 (thoughts and deeds before the performance) indicated that participants under planned conditions were able to set their goals slightly better than under unplanned conditions, but these mean differences were not statistically significant. Only Q6 demonstrated a significant difference between planned (mean=2.00) and unplanned (2.59), suggesting participants found it easier to produce ideas from memories/experience under unplanned rather than planned conditions (see Table 7 in Appendix 2).

The results in Part 3 revealed how they used planning time. Participants were not very conscious about time during planning (Q1–2), and they wrote down the main points to make rather than what to talk about on each element of the prompt card (Q3–5). They were

more likely to plan words and expressions in linguistic planning rather than grammatical structures (Q6–7). A third of the participants did not plan the organization of their talk either on paper or in their mind (Q11–12). It is worth noting that only five of the thirty-two participants thought of what their partners might say, and they (N=4, 1 missing) did not really think about how to answer if their partners said what they had thought they would say (see Q 13–14 in Table 8, Appendix 2).

Regarding perceptions while speaking (Part 4), although there were no significant differences between the two conditions, it is worth noting an overall, counter-intuitive trend in their slightly better perceptions toward unplanned performance (see Table 9 in Appendix 2).

4 Conversation Analysis

The repeated listening and transcribing procedure revealed several distinctive characteristics in the test-takers' interactional patterns between the two conditions that could help explain or build on the above findings. We considered it appropriate to select a dyad whose average turn-length was close to the mean for the whole group (planned=10.02, unplanned=7.97; see Table 4) for the presentation in our analysis with the hope of illustrating interactions that were typical of the participating students. Thus, we selected pair 1 (*S01 & S02*), which indicated the closest values for turn-length under the planned (9.43) and unplanned conditions (7.38). We have used excerpts from this dyad in the following, rather than including those from various dyads, so as to illustrate discourse moves and the co-construction process for the whole test session. The following focuses on

two major characteristics that we identified in the various aspects in our analysis: 1) collaborative interaction without planning time and 2) parallel and asymmetrical interaction with planning. These interactional classifications are those Galaczi (2008) identified in her FCE study.

a Collaborative interaction under unplanned conditions

Frequent short-turn exchanges at beginning. First, the dyads without planning time, tended to develop their thoughts and interactions gradually and collaboratively as they took frequent short turns. Their turns tended to be very short particularly at the beginning of conversations. As they provided initial ideas and listened to their partner's opinions, they gradually started to establish a common understanding of the topic and came up with more ideas that required longer turns to describe.

Excerpt (1) was an initial part of *S01* and *S02*'s unplanned interaction with Task A (happiness). They exchanged their opinions about what aspects of life (e.g., family or friends) would be important for happiness.

Excerpt (1) S01 & S02 (unplanned/Task A)
1 S02: What is:: important thing, do you think?
2→S01: I think (1.0) this one ((pointing out the photo)), talking with friends is
3 S02: Uh
4→S01: the most important (1.0) to be happiness
5 S02: Uh
6→S01: I think.
7 S02: OK. Why do you think so?
8→S01: Becau::se uh when I (.) talk with friends, and hang out with friends, I feel really ha- happy
9 S02: Yeah
10 S01: But money <u>money</u> is also important for us.

This example typically shows how the participants started their conversations without planning time. In response to *S02*'s initiation of the topic, *S01* just identified the most important thing in life, *"talking with friends"*, (lines 2, 4, and 6) without providing any further explanations. This made *S02* ask *S01* the reason for his choice, and *S01* accordingly tried to elaborate on his idea (line 8). As observed in *S01*'s intra-turn pauses and stretching the word *"Becau::se"* (lines 2, 4, and 8), it seems that *S01* was planning on what to say while talking, which made his speech less fluent. After *S02*'s short response token (line 9), *"Yeah"*, *S02* did not offer his own opinion about the topic. Instead, he moved onto another topic (*"money"*), where *S01* this time tried to elicit *S02*'s opinions.

Gradual co-construction in middle. Despite this somewhat awkward start, they gradually developed their ideas in the turns that followed, and simultaneously started to exhibit more engagement in their partner's utterances. After they had exchanged their opinions about *"friends", "money",* and *"love"*, they began to talk about how important having a *"house"* was to achieve a happy life.

Excerpt (2) S01 & S02 (unplanned/Task A)

1 S01: Uh I think house is not impo(h)rtant for me.
2 S02: Oh really?
3 S01: Yes
4 S02: Why do you think so?
5 S01: Because (.) now ah I live in really (.) poor hou(h)se huh huh but I feel much happiness happy,
6 S02: uh
7→S01: so I think this is not important <u>for me.</u>
8→S02: Yeah, I didn't also comment uh:: (1.0) if we don't have good house, uh: maybe ok, because
9 if we have good friends
10 S01: Yes [yes that's right.
11 S02: [Yeah

In response to *S01*'s opinion about the house (line 1), *S02* expressed a little surprise and

asked the reason for this; *"Oh really?", "Why do you think so?"*. This time, the question appeared to be derived from a genuine interest rather than just trying to elicit the partner's opinion to keep the conversation going as in Excerpt (1). Then, *S01* provided a justification for his opinion (lines 5 and 7), with which *S02* agreed. To do so, *S02* elaborated on *S01*'s idea by referring to *S01*'s previous utterance on the importance of friendship. This was further approved by *S01*. It is interesting to note that *S02* added his opinion, following *S01*'s emphasis on *"for me"* in line 7, which was implicitly inviting *S02*'s comment on the topic, by implying that what *S01* said might not apply to *S02*. Excerpt (2) therefore demonstrates that the pair, after having exchanged some initial ideas, started to co-construct their dialogue by engaging more in each other's ideas and elaborating on each other's opinions.

Further collaboration at end. This collaborative tendency was also further observed in the rest of the interaction. After talking about the *"house"*, they switched to the importance of a *"vacation"*.

Excerpt (3) S01 & S02 (unplanned/Task A)

1 S02: How about vacation?
2 S01: Vacation is (1.5) vaca(hhh)tion is (1.0) fun,
3 S02: Yeah
4 S01: but it's just fun
5 S02: uh
6 (2.0)
7 S01: I think (4.0) va- vacation is good for me [because (1.0)
8 S02: [Uh uh
9 S01: it's it's really fu(h)n and (1.5)
10 S02: Uh
11 S01: I can get good experi[ence and (2.0) uh
12 S02: [uh
13→S02: Uh that's right. I think that vacation connected to: this picture= ((showing the friends picture))
14 S01: =A[h yes.
15 S02: [because if we have friend, [we can go: (1.0) uh this this place,

16 S01: [uh huh
17 S02: ah (.5) a:s sightseeing with friends
18 S01: Uh
19 S02: So I think uh:: friends is very very important thing, the best thing
20 S01: <I ((nodding)) totally agree with you> uh but if you (.) we want to go to sta- ah vacation,
21 we: need money, ((pointing out the picture))

In Excerpt (3), *S01* started off the topic by giving a rather negative opinion about the importance of a vacation, *"but it's just fun",* indicating that a vacation is enjoyable but the enjoyment could be of a superficial nature. Then, in line 13, *S02* expressed that they could turn a vacation experience into a more positive one by having good friends accompany them, again referring back to the importance of friends for life.

b Parallel and asymmetrical interaction under planned conditions

Productive start. In contrast to the unplanned interactions with frequent exchanges of short turns, *S01* and *S02* tended to produce longer utterances from the beginning under the planned conditions. Excerpt (4) demonstrates the beginning of planned interactions by the same dyad with Task B (profession), which required them to discuss how difficult it was to be successful in a given list of professions.

Excerpt (4) S01 & S02 (planned/Task B)
1 S01: Which jobs (1.0) is the most difficult?
2→ S02: Uh:: most difficult uh:: I think all picture: have of course uh:: difficulties yeah, but
3 the most difficult (.) job for me (.5) is %I think% this picture (1.0) ((pointing at
4 the painter picture)) is so difficult (.5) to get to the top.
5 S01: Why do you think so?
6→ S02: Ah .hh (.5) I think uh (4.0) %uh five minutes% the first is uh: (.5) many people can buy
7 this picture or not, this is the uh (.5) the most difficult thing. (1.0) Maybe before before the
8 person buy the picture, this person is so poor.
9 S01: Uh[:
10→S02: [Uh So this is so difficult job. How about you?
11→S01: I basically: agree with you, but but singer (.5) and soccer player [(.5) are also (.5) difficult
12 S02: [uh uh
13 S01: to be, because they if you want to be like them, ah (1.0) you need (1.5) an talent

Initiated by *S01*'s question, *S02* explained that a painter would be the most difficult profession (lines 2–4). His utterance was very slow and not very sophisticated, but compared to the beginning of their unplanned interaction (Excerpt (1)), it is apparent that he attempted to elaborate more on his idea. Then, in response to *S01*'s request for the reason, *S02* produced a longer turn (lines 6–8).

After justifying why becoming a successful painter could be difficult, *S02* reiterated his opinion, and said *"How about you?"* in line 10, as if to indicate that it was now *S01*'s turn to present what *S01* had prepared to talk about during the planning time. Although *S01* initially said, "*I basically: agree with you"* following the question, he did not comment further on *S02*'s previous talk, and he simply started to talk about how difficult becoming a singer and a football player could be and why he thought so. Both *S01* and *S02* successfully used a complex sentence in this initial part of the task (lines 7–8 and 13). It was also noted that the articulation rate in these planned, longer turns was generally slower than that of unplanned short turns.

Stagnant middle. Despite a parallel but productive start, a stagnant period soon followed. They attempted to exchange their opinions about business people in Excerpt (5). However, it appears that they ran out of ideas and failed to develop interaction as indicated by the number of filled and unfilled pauses (lines 3, 5, 7, 10, 11, and 13) and repetitions (lines 2, 12, and 14).

Excerpt (5) S01 & S02 (planned/Task B)
1 S02: Then how about business man?
2→ S01: Business man

3 (2.5)
4 S02: Compared to the other pictures, this is not difficult to get to the top.
5 (.5)
6 S01: Uh
7 (1.5)
8 S01: If you wan- if you want to be: (.5) business man, you need talent?
9 S02: uh
10 (.5)
11 S02: Uh:: (.5) we need effort= we need, most need thing, I think.
12→S01: =Effort
13 (2.0)
14→S01: Effort (.5) is impor(h)tant huh [huh
15 S02: [Yeah

Asymmetrical end. *S02* occasionally attempted to produce long utterances in the final part of this dialogue, as in Excerpt (6), while *S01* tended to be rather passive by mostly giving response tokens only.

Excerpt (6) S01 & S02 (planned/Task B)

1 S02: Ah I also (1.5) doctor is (.5) not difficult to get to the top, because (.5) in doctor, there
2 are a lot of sick people. If there are many sick people, ah the person (.) ca(h)n wo(h)rk.
3 S01: Ye(h)s
4 S02: Yes
5 S01: That's (2.0) need (.) need much money, I thi(h)[nk and intelligence
6 S02: [ah ah and we often
7 need doctor. (1.0) We need sports player >because many people like sports<
8 and we need singer. We often listen to the music,
9 S01: Uh huh
10 S02: but this picture, some people need picture,
11 S01: Uh
12 S02: but in my opinion, we don't need picture. (.5) yeah so this is so difficult to get to the top
13 S01: Uh I think pictures are (1.5) an entertainment for (.5) rich rich people
14 S02: Uh
15 S01: So (2.0) I think (.5) the painter is the most difficult to get the top.
16 S02: Yeah, I agree.

In addition to *S02*'s sequential and quantitative dominance in this part, *S01* failed to contribute to the interaction, which made the interaction asymmetrical. This is in sharp contrast to their effective and cooperative exchanges in the latter part of the unplanned interaction.

Figure 2 describes the trajectories of turn-lengths undertaken by *S01* and *S02* to help to better understand the overall trends. Each started saying rather a limited number of words under the unplanned conditions, but they then gradually increased turn-lengths (except for that in the eighth turn). In addition, although their turn-length was rather asymmetrical at the beginning, they tended to have balanced turn-lengths from the middle of the interaction (i.e., the thirteenth turn).

In contrast, the planned interaction was characterized by a limited number of turns and there were huge discrepancies in turn-length between the two speakers and across turns. Unlike their unplanned interaction, they produced long turns at the beginning of the interaction (first and second turns by *S01* and fourth turn by *S02*), but then the turn-lengths gradually decreased to reach the bottom in the middle between the ninth and eleventh turns. Although the turn lengths then suddenly increased from time to time (e.g., in the twentieth turn for *S02* and the fifteenth for *S01*) they soon declined into shorter turns.

Add Figure 2 around here

IV Discussion

The various methods of analysis have provided different and supplementary pictures of planned and unplanned performances. As in the previous planning studies on testing (e.g., Wigglesworth & Elder, 2010), the present study suggested complex relationships between test scores and discourse analytic measures. Score analysis revealed that pre-task planning slightly upgraded test-takers' speech in terms of fluency and complexity, and discourse analytic measures suggested improvements in the breakdown of fluency and longer turn

length under planned conditions. Discourse analytic measures also showed that the planned conditions were detrimental to the speed of fluency. The analysis of the questionnaire indicated that test-takers did not seem to have used the given planning time very strategically according to the trend in limited planning effects, and they even felt it easier to produce their speech under unplanned conditions.

The CA data offered useful insights to understand these findings. The decreased speed of fluency under the planned conditions could be related to their attempts to produce longer utterances under these conditions. When the test-takers attempted to produce longer turns while recalling what they had planned, their intra-turn speech rate tended to slow down. In contrast, more animated, spontaneous shorter turns under the unplanned conditions enabled them to talk more quickly at times, although the increased cognitive demands in relation to on-line planning under the unplanned conditions seemed to make their interactions contain more pauses especially at the beginning. As they needed to gather their thoughts while planning on-line, one tended to play an interviewer's role, gradually eliciting opinions from his/her partner. Thus, the part at the beginning under unplanned conditions can be characterized as a series of short turns with breakdown features.

In contrast, the beginning of planned interactions was characterized by longer turn-lengths, as test-takers presented what they planned to say during the planning stage. This might explain the results from score analysis, where a slight but significant increase in complexity scores was observed under the planned conditions, although this did not make a significant difference to the syntactic complexity of the discourse analytic measure. The beginning of the planned interaction was productive but was characterized as a parallel

pattern, where both parties contributed to the conversation but their interactions were not mutually developed. Their utterances tended to be longer but resembled a series of monologues, often connected with the mechanical use of *"How about you?"* as seen in line 10 in Excerpt (4). It is worth pointing out that such unnatural turn-taking was also observed in Van Moere's (2007) study of group oral tests with 1 minute of planning time.

Pre-task planning made them prepare for their speech, and as such, enabled them to produce longer turns on average. However, the analysis of interactional data also substantiated that the planned ideas and language seemed to become exhausted and test-takers soon fell into a stagnant period. This stagnation under the planned conditions seemed to be reflected in the counter-intuitive results from the questionnaire where the participants found it more difficult to generate ideas under planned conditions than unplanned conditions (Z=-2.301, p=0.021).

In contrast, a lack of planning led to a gradual increase in turn-length. The test-takers without planning time seemed to engage in collaborative mode as they spoke. As illustrated in Excerpt (2), they incorporated their and their partner's prior talk into the current topic to develop interactions collaboratively and coherently. Thus, the conversational path was not individually paved but was mutually developed through their exchange of turns.

In the final part of planned interactions, the turn-taking pattern appeared to be clumsy, which was in great contrast to the gradually developing trend under unplanned conditions. Rather than attempting to develop ideas together, they just seemed busy expressing their own ideas in turns.

V Conclusion

This study employed more fine-grained process-oriented analysis in recognizing that exclusive reliance on summative approaches to investigating pre-task planning cannot provide a comprehensive account of its impact on performance, and it considered participants' opinions to understand the complex relationships between pre-task planning and performance.

The study identified the possibility that planning conditions affect the mode of discourse in paired tests. Their interactions with planning time, particularly at the beginning of an interaction, resembled a series of monologues rather than a dialogue. Not only did they produce longer turns, they seemed to concentrate more on delivering what they had prepared during the planning time, expressing little interest in what their partner had said. As a result, unlike the unplanned interactions where the test-takers more cooperatively approached the task, they made fewer attempts in developing a topic initiated by their partner or in incorporating their partner's ideas into their own speech when expanding the discourse. Thus, the individualistic approach to the task under the planned conditions led to a parallel pattern of interaction. This was in sharp contrast to a more collaborative pattern of interaction, which test designers originally intended to elicit by using such paired speaking formats as those in this study.

The findings of the present study have provided several implications for classroom teaching and language testing. The possibility of changing discourse modes in a classroom context by using planning time makes teachers aware of using planning according to their aims in teaching. For example, a dialogic task without planning time seems more effective

when it is aimed at developing interactional competence (Young, 2000).

Implementing pre-task planning-time prior to a paired format in the practice of language testing may not be advisable. Although providing a 3-minute planning time seemed to slightly benefit test-takers, this study identified a concern that planning time might change interactional patterns elicited in paired tests from a more collaborative discourse to a more parallel discourse. Since a paired test format aims to measure the extent to which candidates can effectively communicate by interacting with each other, this discoursal change functions against tapping into the construct that this format should actually be measuring. Considerations should therefore be given to not providing pre-task planning times that could adversely affect the construct validity of the test to assess interactional competence in paired speaking tests as test designers have intended.

In addition, it might be worthwhile for examination boards to reconsider the test duration for paired tasks. The CA data revealed that collaborative interaction was gradually co-constructed under unplanned conditions over the given 5-minute period. Consideration should be given as to whether a 3-minute performance time, as is currently applied in some standardized tests such as FCE, can provide sufficient time to fully assess students' interactional abilities.

As was explained earlier, despite TBLT and language testing having a reciprocal relationship, the previous findings on pre-task planning by each strand have not always been effectively connected. It is hoped that the multifaceted approach presented in this study will open up new avenues to understanding complex relationships between pre-task

planning and interaction by means of bringing in participants' opinions and their dynamic co-constructing processes in interactions.

VI Acknowledgements

This study was supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan [grant no. 21720207]. We are grateful to Dr Parvaneh Tavakoli for her valuable comments on an earlier draft of this paper. We also gratefully acknowledge the insightful comments made by the Language Testing reviewers that helped improve this paper.

VII References

Atkinson, J. M. & Heritage, J. (Eds.) (1984). *Structures of social action: Studies in Conversation Analysis.* Cambridge, New York: Cambridge University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*, 1–25.

Cambridge ESOL (2008). *Cambridge First Certificate in English 1 for updated exam Student's Book with answers: Official Examination papers from University of Cambridge ESOL Examinations (FCE Practice Tests)*. Cambridge: Cambridge University Press.

Cambridge ESOL (2012). *Cambridge English First: First Certificate in English (FCE): Handbook for teachers*, accessed on 06/12/2012 at https://www.teachers.cambridgeesol.org/ts/digitalAssets/117578_Cambridge_English_Fi

rst_FCE_Handbook.pdf

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Duff, P. (1986). Another look at interlanguage talk: Taking task to task. In R. Day (Ed.), *Talking to learn: conversation in second language acquisition* (pp. 147-181). Rowley, MA: Newbury House.

Elder, C. A., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*, 347–368.

Elder, C. & Iwashita, N. (2005). Planning for test performance: What difference does it make? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219–238). Amsterdam: John Benjamins.

Ellis, R. (2005). Planning in language testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 217–218). Amsterdam: John Benjamins.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production, *Applied Linguistics*, *19*, 474-509.

Educational Testing Service (2012). TOEFL IPT: CEFR Mapping Study, online, accessed on 03/12/2012 at: http://www.ets.org/toefl_itp/research/

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking* (pp.65-111). Cambridge: Cambridge University Press.

Foster, P. & Skehan, P. (1996). The influence of planning and task-type on second language performance. *Studies in Second Language Acquisition, 18,* 299–323.

Foster, P. Tonkyn, A. & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21,* 354–375.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*, 89-119.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*, 585-402.

Iwashita, N., Elder, C. & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 21,* 401–436.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.* Cambridge: Cambridge University Press.

McCarthy, M. (2010). Spoken fluency revisited, *English Profile Journal, 1 (1),* online, accessed on 01/05/2012
at: http://journals.cambridge.org/action/displayIssue?jid=EPJ&volumeId=1&seriesId=0&issueId=01

McCarthy, P.M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity Assessment. *Behavior Research Methods. 42,* 381-392.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83–108.

Myford, C. M. & Wolfe, E. W. (2003) Detecting and measuring rater effects using many-facet Rasch measurement: part 1. *Journal of Educational Measurement, 4(4),* 386-422.

Nakatsuhara, F. (2011). Effects of the number of participants on group oral test performance, *Language Testing, 28(4):* 483-508.

O'Sullivan, B. (2008). *Modelling performance in oral language testing*. Frankfurt: Peter Lang.

Samuda, V. & Bygate, M. (2008). *Tasks in second language learning.* Basingstoke: Palgrave.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193–216). Amsterdam: John Benjamins.

Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.

Taylor, L. (Ed.) (2011). *Examining Spe*aking*: Research and practice in assessing second language speaking.* Cambridge: UCLES/Cambridge University Press.

Van Moere, A. (2007). *Group oral test: How does task affect candidate performance and test score?* Unpublished PhD thesis, Lancaster University.

Weir, C. J. (2005). *Language testing and validation: Evidence-based approach.* London: Palgrave Macmillan.

Weir, C. J., O'Sullivan, B. & Horai, T. (2006). Exploring Difficulty in Speaking Tasks: an Intra-task Perspective. *IELTS Research Report No.6* (pp. 119-160) British Council and IDP Australia.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*, 101–122.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Task based learning* (pp. 186–209). London: Addison Wesley Longman.

Wigglesworth, G. & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*, 1-24.

Wright, B. and Linacre, M. (1994). *Reasonable mean-square fit values*, accessed at 04/12/2012 at http://www.rasch.org.

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22*, 463–508.

Young, R. (2000). *Interactional competence: challenges for validity.* Paper presented at a joint symposium on 'Interdisciplinary interface with language testing', held at the Annual Meeting of the American Association for Applied Linguistics and the Language Testing Research Colloquium, 11/03/2000, Vancouver, British Columbia, Canada, on-line, accessed on 01/05/2012 at: http://www.wisc.edu/english/rfyoung/IC_C4V.Paper.PDF

Appendix 1: Rating scales (Modified from Iwashita et al, 2001)

Fluency

8	Speaks fairly fluently with only occasional hesitation, false starts and modification of attempted utterance. Speech is only slightly slower than that of a native speaker.
7	
6	Speaks more slowly than a native speaker due to hesitations and word-finding delays.
5	
4	A marked degree of hesitation due to word-finding delays or inability to phrase utterances easily.
3	
2	Speech is quite disfluent due to frequent and lengthy hesitations or false starts.
1	
0	Speech is so halting and fragmentary that conversation is impossible.

Accuracy

8	Errors are not unusual, but rarely major.
7	
6	Manages most common forms, with occasional errors, major errors present.
5	
4	Limited linguistic control: major errors frequent.
3	
2	Clear lack of linguistic control even of basic forms.
1	
0	No linguistic control even of the most basic forms.

Complexity

8	Attempts a variety of verb forms (e.g. passives, modals, tense and aspect), even if the use is not always correct. Takes risks grammatically in the service of expressing complex meaning. Regularly attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is awkward or incorrect.
7	
6	Mostly relies on simple verb forms, with some attempt to use a greater variety of forms (e.g. passives, modals, more varied tense and aspect). Some attempt to use coordination and subordination to convey ideas that cannot be expressed in a single clause.
5	
4	Produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty.
3	
2	Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect ideas across clauses.
1	
0	No awareness of basic grammatical means.

Appendix 2. Questionnaire results

Table 5: Part 1. About tasks (task comparison)

Item	Mean				Kruskal-Wallis Test
	Happiness (N=16)	**Profession** (N=16)	**Café** (N=16)	**Tourists** (N=16)	
Q1 Vocabulary was: *1. very difficult – 5. very easy*	4.50	4.13	3.75	4.13	χ^2=4.054 *p*=.256
Q2 Grammar was: *1. very difficult – 5. very easy*	4.44	4.00	3.81	3.94	χ^2=4.069 *p*=.254
Q3 Topic was: *not familiar at all – 5. very familiar*	3.63	3.44	3.31	2.69	χ^2=5.077 *p*=.166
Q4 Information was: *1. very abstract – 5. very concrete*	3.44	3.69	3.94	3.00	χ^2=5.621 *p*=.132
Q5 Preparation time was: *1. too short – 3. Appropriate – 5. too long*	(N=8) 2.13	(N=8) 2.75	(N=7) 2.86	(N=8) 2.38	χ^2=4.362 *p*=.225
Q6 Task time was long: *1. too short – 3. Appropriate - 5. too long*	2.44	2.63	2.25	2.38	χ^2=1.486 *p*=.685

Table 6: Part 1. About tasks (comparison of planned and unplanned conditions)

Item	Mean		Wilcoxon Signed Rank Test
	Planned (N=32)	**Unplanned** (N=32)	
Q1 Vocabulary was: *1. very difficult – 5. very easy*	4.06	4.19	Z=-.883 *p*=.377
Q2 Grammar was: *1. very difficult – 5. very easy*	4.09	4.00	Z=-.322 *p*=.748
Q3 Topic was: *1. not familiar at all – 5. very familiar*	3.19	3.34	Z=-.493 *p*=.622
Q4 Information was: *1. very abstract – 5. very concrete*	3.47	3.56	Z=-.312 *p*=.755
Q5 Preparation time was: *1. too short – 3. Appropriate – 5. too long*	2.52	N/A	N/A
Q6 Task time was long: *1. too short – 3. Appropriate - 5. too long*	2.44	2.41	Z=-.179 *p*=.858

Table 7: Part 2. What I thought of or did before I started

	Item	Mean (SD)		Wilcoxon Signed Rank Test
		Planned	**Unplanned**	
Goal setting	**Q1** I read the task very carefully to understand what was required	4.16 (.92)	4.03 (.90)	Z=-.655 *p*=.512
	Q2 I thought of how to provide my ideas to respond well to the topic	3.88 (1.21)	3.50 (.92)	Z=-1.692 *p*=.091
	Q3 I thought of how to convey my message to my partner clearly	3.84 (1.11)	3.63 (1.10)	Z=-1.072 *p*=.284
	Q4 I understood the instructions for this	4.38 (.75)	4.56 (.56)	Z=-1.213

	speaking task completely			*p*=.225
Generating ideas	**Q5** I had enough ideas to speak about this topic	2.67 (1.10)	2.67 (1.09)	Z=-.587 *p*=.557
	Q6 I felt easy to produce enough ideas for the interaction from memory/experience	2.00 (.84)	2.59 (1.13)	**Z=-2.301** ***p*=.021**
	Q7 I know a lot about this type of speaking task i.e., how to interact in pairs	1.97 (.97)	2.00 (.92)	Z=-.021 *p*=.983

1: Strongly disagree – 5: Strongly agree

Table 8: Part 3. What I thought of or did in planning stage

	Item	**Mean (SD)**
Time element	**Q1** I thought of most of my ideas for the interaction before planning how to deliver them in the interaction	3.44 (1.16)
	Q2 During the period allowed for planning, I was conscious of the time i.e., how to use the planning time/how much time is left	1.90 (1.09)
Task specific planning	**Q3** I thought of what to talk about for all elements of the prompt card	3.03 (1.28)
	Q4 I thought of which one or two elements I would eventually like to choose in the decision making phase.	3.29 (1.32)
	Q5 I wrote down the points I wanted to make based on the visual information in the prompt card	4.03 (.86)
Linguistic planning	**Q6** I wrote down the words and expressions I needed to fulfil the task	3.19 (1.40)
	Q7 I wrote down the grammatical structures I need to fulfil the task	1.57 (.86)
Language used when planning	**Q8-10** I took notes only in English, only in Japanese or in both	English: 11 (34.4%) Japanese: 9 (28.1%) Both: 11 (34.4%) Neither: 1 (3.1%)
Organization	**Q11-12** I planned how to organize my talk on paper or in mind or both before starting to speak	Paper: 3 (9.4%) Mind: 9 (28.1%) Both: 8 (25.0%) Neither: 12 (37.5%)
Interaction planning & practicing	**Q13** I thought of what my partner might say about each element in the prompt	Yes: 5 (15.6%) No: 27 (84.4%)
	Q14 If yes, I planned how to answer my partner, if he/she says what I thought he/she would say (N=4)	2.75 (.96)
	Q15 After finishing my planning, I practised what I was going to say in my mind until it was time to start	2.22 (.97)

1: Strongly disagree – 5: Strongly agree

Table 9: Part 4. What I thought of or did while I was speaking

	Item	**Mean (SD)**		**Wilcoxon Signed Rank Test**
		Planned	**Unplanned**	
Idea development (ability) & completing the	**Q1** I felt it was easy to give my opinions during the interaction	2.28 (1.08)	2.35 (1.11)	Z=-.577 *p*=.564
	Q2 I was able to express my ideas using suitable words	2.41 (.911)	2.35 (.80)	Z=-.714 *p*=.475

task	**Q3** I was able to express my ideas using correct grammar	1.97 (.93)	2.00 (.86)	Z=-.243 *p*=.808
	Q4 I was able to put sentences in logical order	2.06 (.95)	1.97 (.80)	Z=-.688 *p*=.491
	Q5 I was able to connect my ideas smoothly in the whole interaction	1.84 (1.04)	2.13 (1.09)	Z=-1.291 *p*=.197
	Q6 I felt it was easy to complete the task	1.84 (.88)	2.06 (1.03)	Z=-1.171 *p*=.242
	Q7 While I was speaking, I used all ideas that I had planned	2.35 (1.08)	N/A	N/A
Monitoring	**Q8** I was listening and checking the correctness of the contents while I was talking	3.34 (1.10)	3.55 (1.03)	Z=-1.090 *p*=.276
	Q9 I was listening and checking the correctness of sentences while I was talking	3.44 (1.13)	3.48 (1.00)	Z=-.486 *p*=.627
	Q10 I was listening and checking whether the words fit the topic while I was talking	3.28 (1.05)	3.48 (1.00)	Z=-1.147 *p*=.251
Interacting with partner	**Q11** When my partner was talking, I was fully concentrating in what he/she was talking about	4.03 (.78)	4.23 (.72)	Z=-1.387 *p*=.166
	Q12 When my partner was talking, I was thinking about what I should say after he/she finishes the talk	3.47 (1.08)	3.42 (.96)	Z=-.089 *p*=.929

1: Strongly disagree – 5: Strongly agree

Appendix 3: Transcription notation (Modified from Atkinson & Heritage, 1984)

Unfilled pauses or gaps	Periods of silence. Micro-pauses (less than .2 second) are shown as (.); longer pauses appear as a time within parentheses. E.g. (.5) represents five tenths of a second.
Colon (:)	A lengthened sound or syllable; more colons prolong the stretch
Dash (-)	A cut off, usually a glottal stop
.hhh	Inhalation
Hhh	Exhalation
hah, huh, heh	Laughter
(h)	Breathiness within a word
Punctuation	Intonation rather than clausal structure; a full stop (.) is falling intonation, a question mark (?) is rising intonation, a comma (,) is continuing intonation
Equal sign (=)	A latched utterance, no interval between utterances
Open bracket ([ )	Beginning of overlapping utterances
Percent signs (% %)	Quiet talk
Asterisks (* *)	Creaky voice
Empty parentheses ( )	Words within parentheses are doubtful or uncertain
Double parentheses (( ))	Non-vocal action, details of scene.
Arrows (><)	The talk speeds up
Arrows (<>)	The talk slows down
<u>Underlining</u>	A word or sound is emphasised
Psk	A lip smack
Tch	A tongue click
Italics	Japanese words
Arrow (→)	A feature of interest to the analyst

Figure 1. Facet map (all facet vertical rulers)

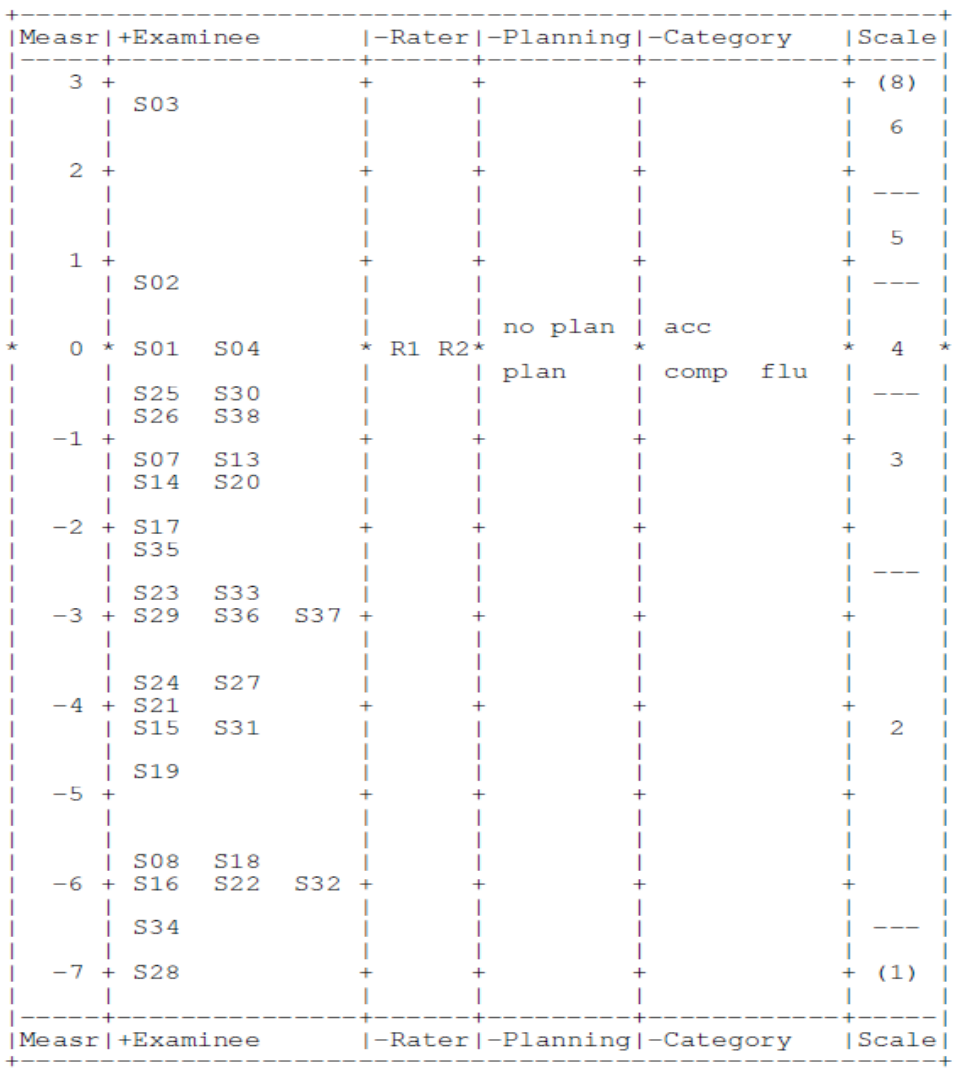

Figure 2. Changes of turn length (S01 & S02)